Winning Space Race with Data Science

<Joshua Lacerda>
<Date: June 2023>

Outline

Executive Summary

- Summary of methodologies
  - Data Collection from SpaceX API and public reports
  - Data wrangling and analysis of potential for successful rocket landings
  - SQL and EDA to evaluate outcomes in relation to multiple parameters
  - Data Visualization in charts and dashboards to facilitate reporting and analysis.
  - Machine Learning to find the best model that predicts the outcome with higher accuracy
- Summary of all results
  - Four models created to predict if SpaceX's First Stage rocket lands successfully or not. We use Logistic Regression; Support Vector Machine; Decision Tree Classifier and K nearest neighbors.
  - All models predicted with similar accuracy of 83.33%. However, the Tree Classifier stands out with a 94.4% accuracy. This is a good prediction, but not ideal. Hopefully we'll keep gathering more data and analysis to improve the accuracy rate.

Introduction

- Project background and context
  - In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. In this lab, you will collect and make sure the data is in the correct format from an API. The following is an example of a successful and launch.
- Problems you want to find answers
  - Be able to predict the likelihood of the first stage of a rocket landing successfully
  - Which parameters influence in the outcome of such success
  - Increase the success rate of first stage landings.

Section 1

Methodology

Methodology

Executive Summary

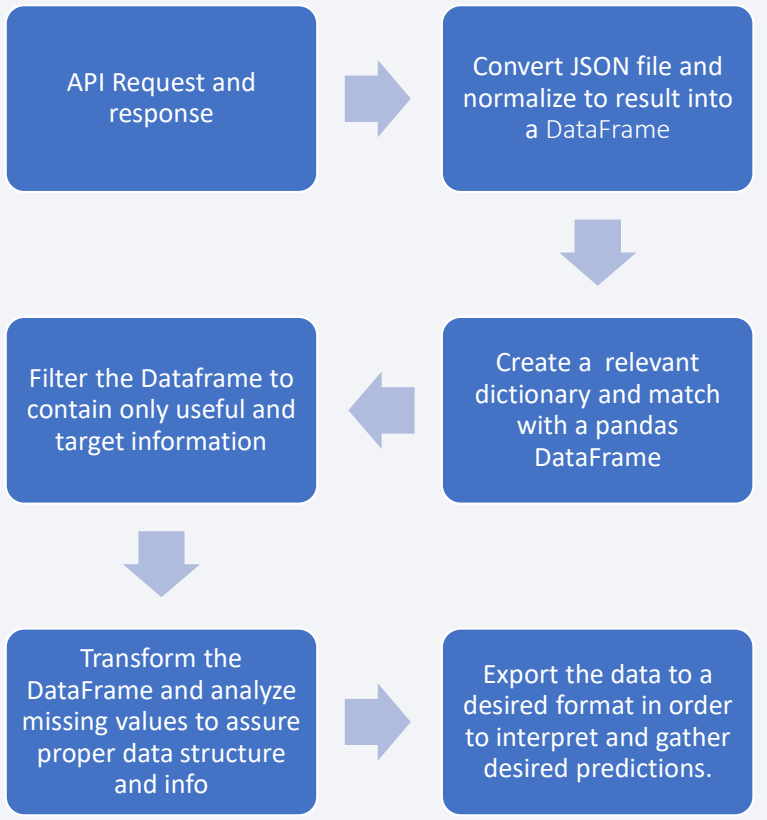
- Data collection methodology:
  - Combination of Data from SpaceX's API and Web scraping
- Perform data wrangling
  - Data transformation. Removal of irrelevant fields and entries.
  - Conversion of landing successes into binary format for future model prediction
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Describe how data sets were collected.
  - From the SpaceX REST API at api.spacexdata.com/v4/launches/past which is the endpoint containing the information from previous launches and landing history.
  - This API provides a JSON file. This file is in a dictionary format, and it makes it possible for us to use and extract pertinent information for our use.
  - The file is than normalized and pushed into a .csv format, which allows us to manipulate with ease.
  - Web scraping Wikipedia from SpaceX's webpage and extracting the html with BeautifulSoup Python package.

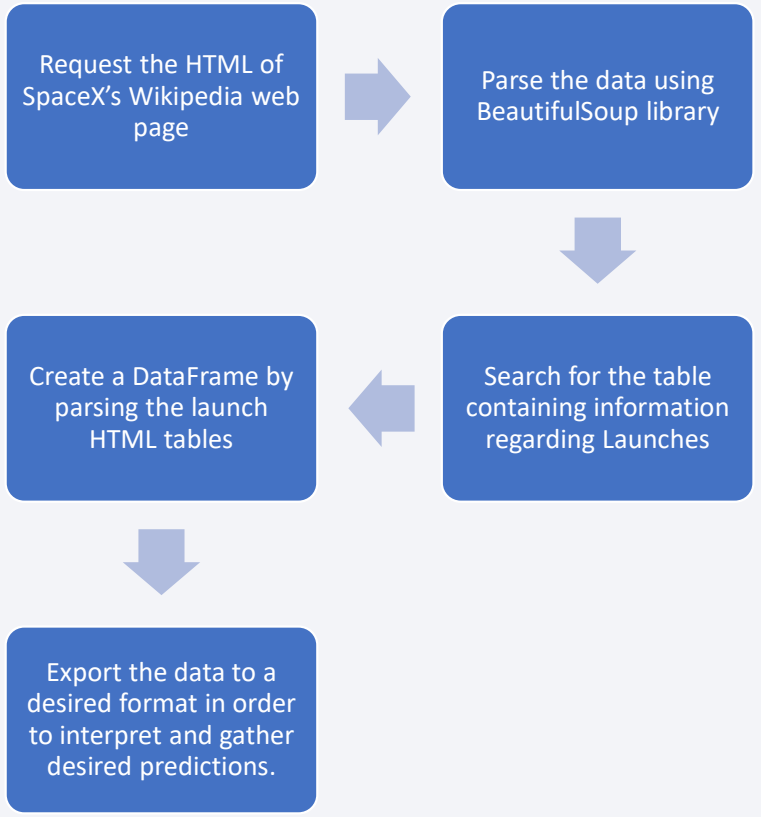
Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts
- GitHub Link – Data Collection



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- GitHub Link – Data Scraping



Data Wrangling

- Describe how data were processed
  - Understand the data and separate the successful landing to unsuccessful ones
  - Successful landings are converted to the value of 1
  - Unsuccessful landing are converted to the value of 0
- You need to present your data wrangling process using key phrases and flowcharts
  - Separate the information as per Launch Sites by using value_count() method.
  - Calculate the number of occurrence of mission outcome per orbit
  - Create a landing outcome label from the Outcome column and classify into 1's and 0's
  - Determine the success rate overall and per Launch site
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose
- GitHub Link – Data Wrangling

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - The Exploratory Data Analysis was intended to identify how the independent variables could affect the outcome of the dependent variable, in this case, successful launches and landing.
  - Charts plotted
    - `sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect = 5)`
    - `sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)`
    - `sns.catplot( y="PayloadMass", x="LaunchSite", hue="Class", data=df, aspect = 5)`
    - `sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)`
    - `sns.barplot(y="Class", x="Orbit", data=df_orbit)`
    - `sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)`
    - `sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)`
    - `sns.lineplot(data = df, x="Date" , y="Success Rate")`
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
- GitHub Link – EDA with Data Visualization

EDA with SQL

- SQL Queries performed
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose
- GitHub Link – EDA with SQL

Build an Interactive Map with Folium

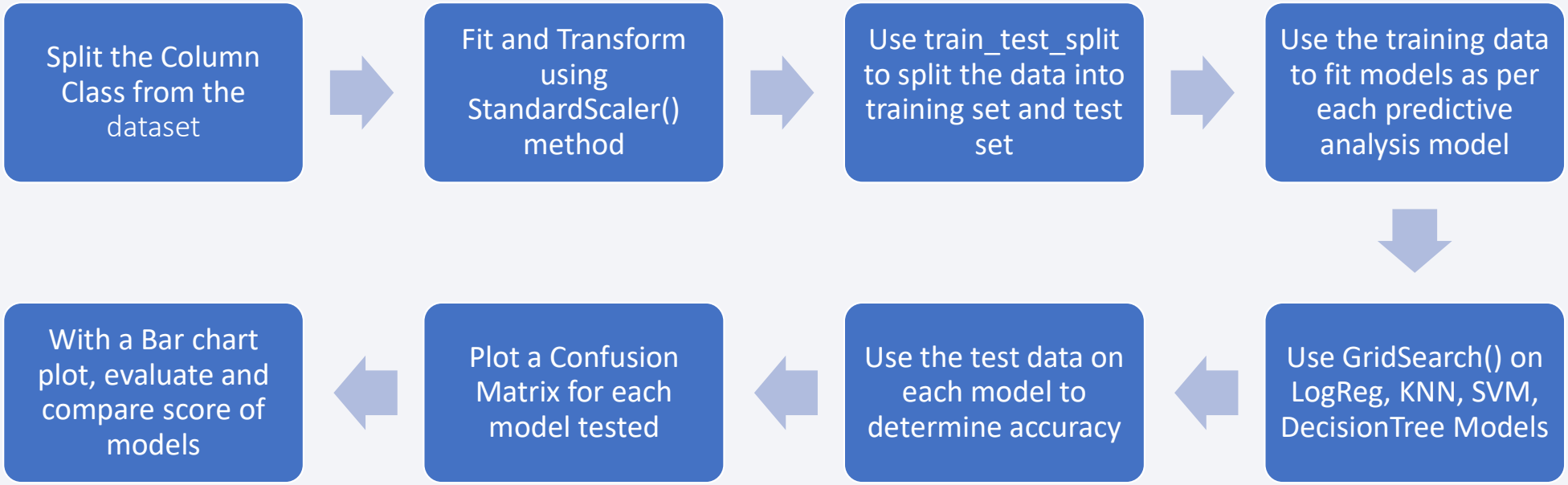
- Multiple Maps were created with the Folium package.
- This package allows us to plot maps of many kinds and complexities. Making maps interactive and meaningful.
- Plotting a map may not mean much without a proper description and legends. And in order to present meaning and insights to a map, there are different tools from which we can select.
- Markers and circles were used to better display the information in our DataFrame and to facilitate the interpretation of the data presented.

- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose
- GitHub Link – Interactive map with Folium

Build a Dashboard with Plotly Dash

- A dashboard was created using Plotly Dash Application. And in the application we generated two interactive charts. One Pie Chart and a second Scatter Plot chart
- The Pie Chart contains information of the success rate of Launches and landings for each site.
- The Scatter plot shows us the success of these Launches and Landings in relation to the Payload Mass (in kg) that the rocket carried
- The idea of an interactive dashboard is such that you can visually analyze the data you're working with and make decisions faster.
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
- GitHub Link – Dash_Interactivity

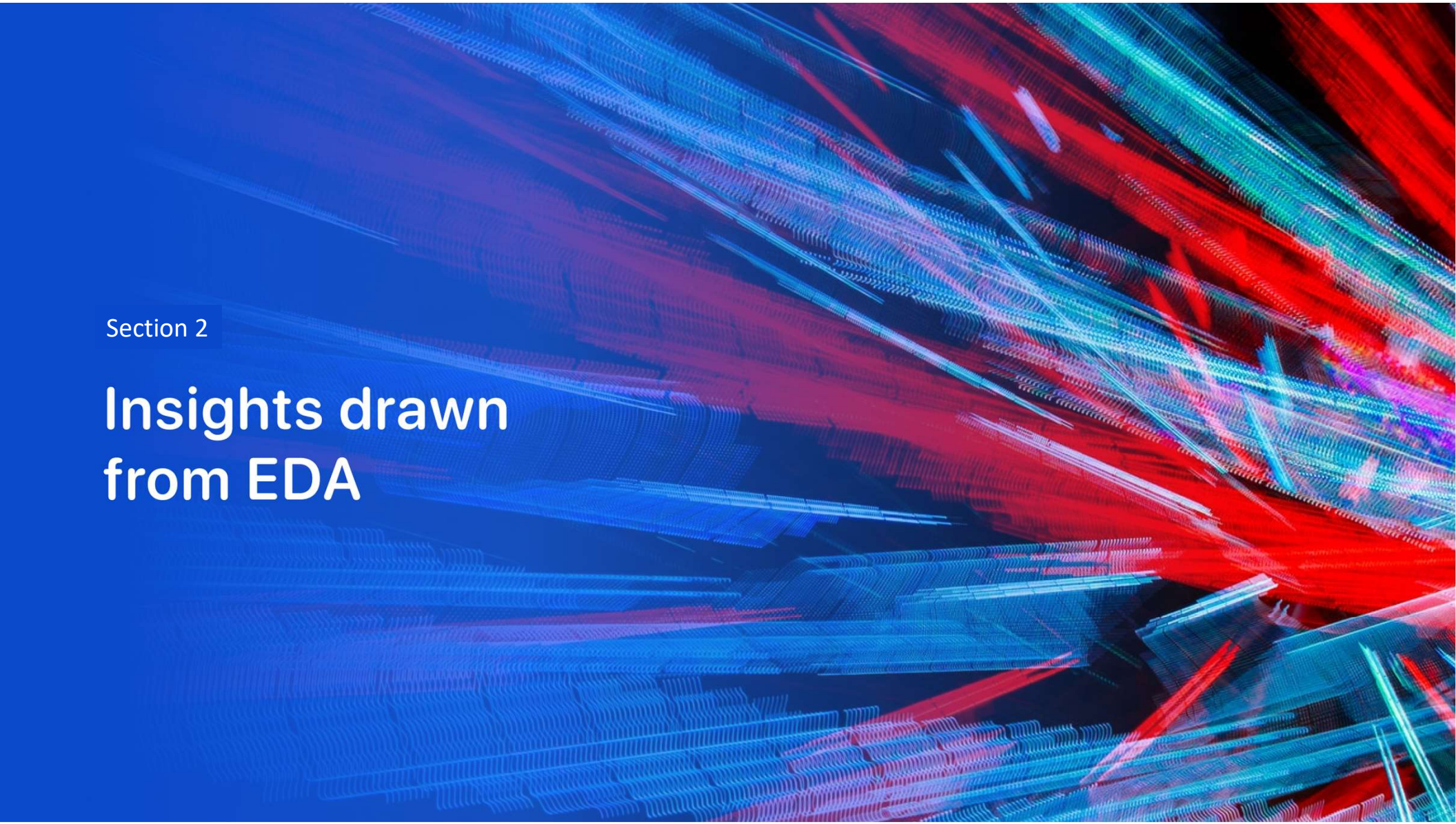
Predictive Analysis (Classification)



- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
- GitHub Link – Predictive Analysis

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

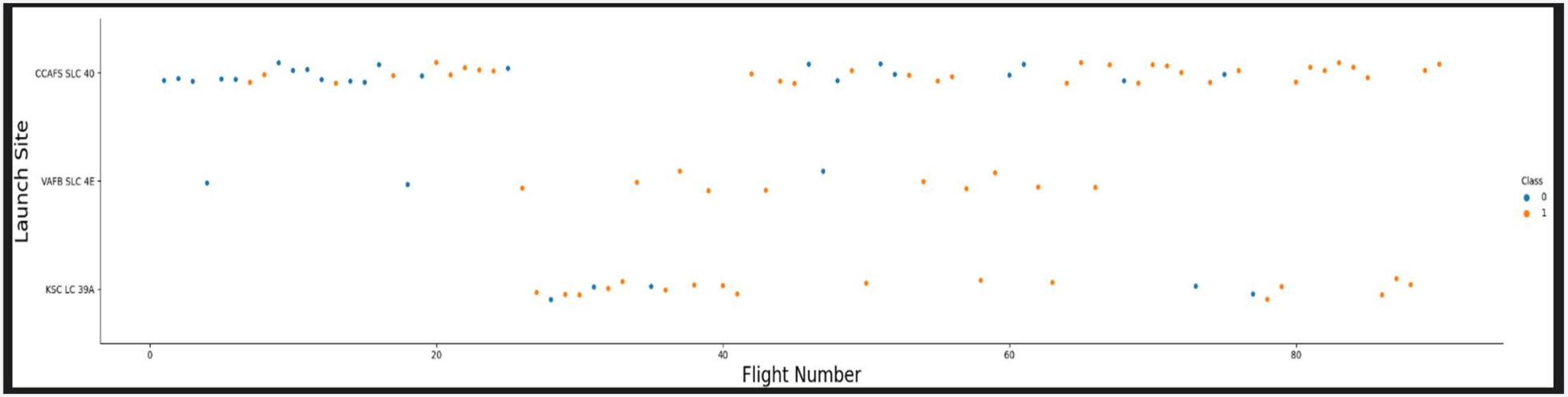


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

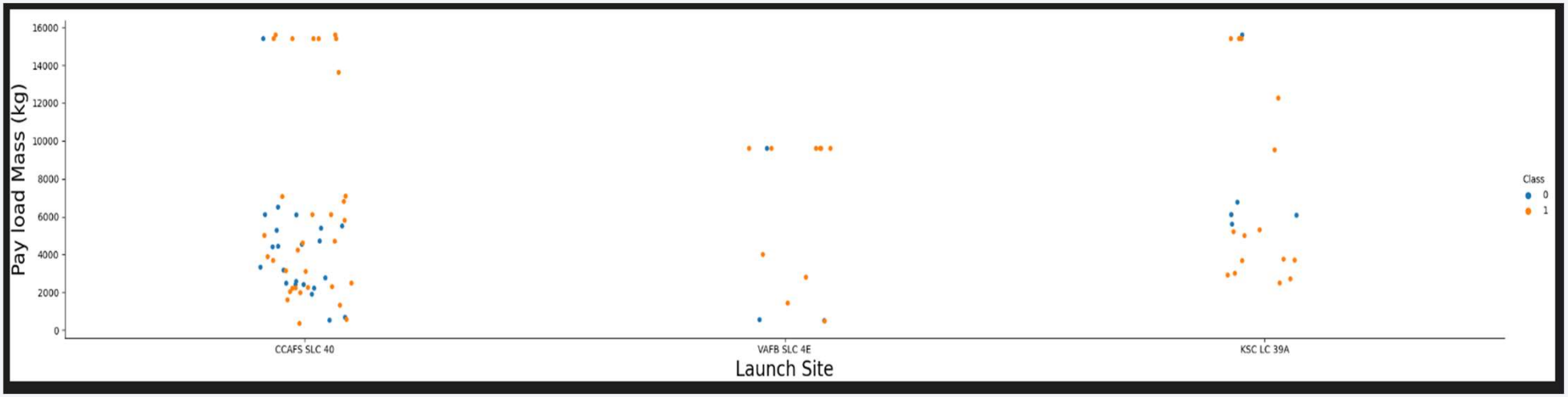
- Show a scatter plot of Flight Number vs. Launch Site



- We compare the flight number and launch site locations and evaluate the success (Class 1) and failures (Class 0) for each mission. The result shows good amount of success from different location, but it is important to highlight that the number of failures are decreasing substantially as the flights numbers increase. Showing a successful progress by each launch.

Payload vs. Launch Site

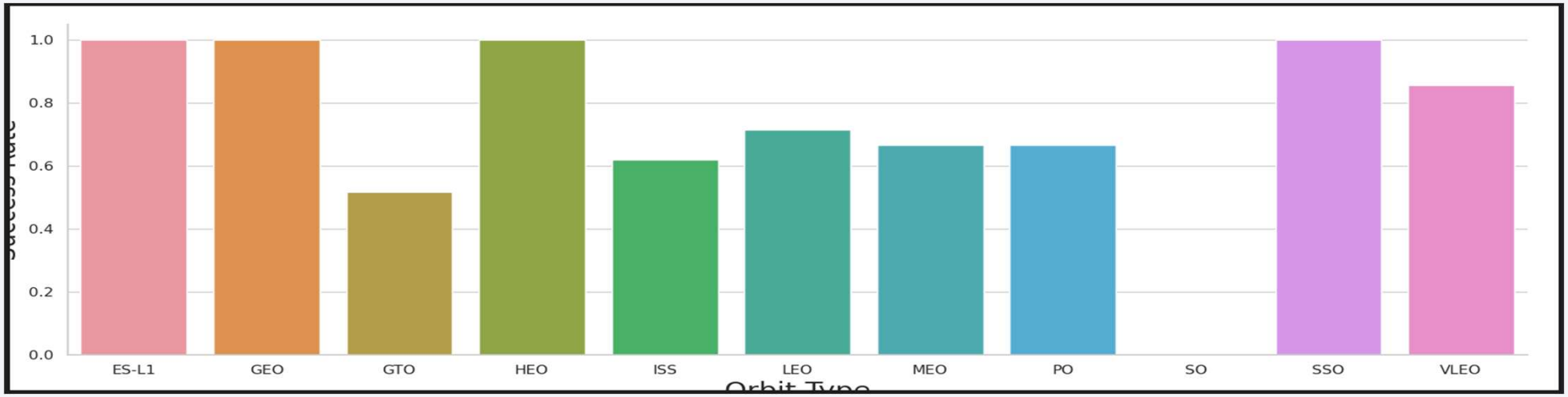
- Show a scatter plot of Payload vs. Launch Site



- Different launch sites present different limitations, successes and failures. As shown, the VAFB site present a limitation of 10000kg of capacity as per CCAFS and KSC sites can launch rockets as heavy as 16000kg.
- We can also observe that the CCAFS launch site is one of the most used out of all launch sites.

Success Rate vs. Orbit Type

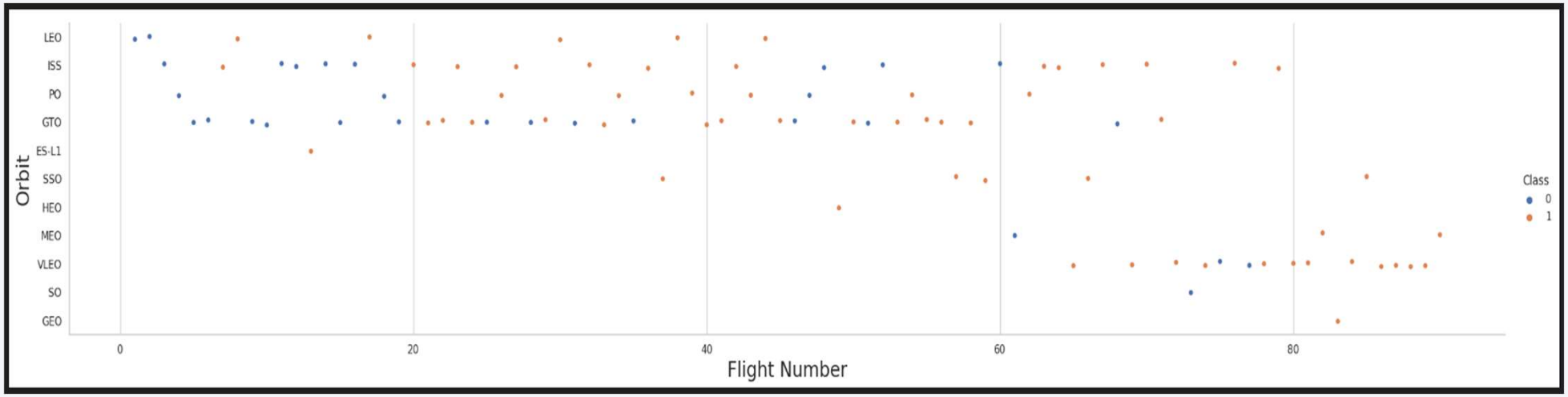
- Show a bar chart for the success rate of each orbit type



- The success rate of the Orbits ES-L1, GEO, HEO and SSO are 100% successful. As for the SO orbit has never had a successful launch.
- The information on this bar chart is missing the number of launches per site, for us to determine how efficient a launch site is or can be.

Flight Number vs. Orbit Type

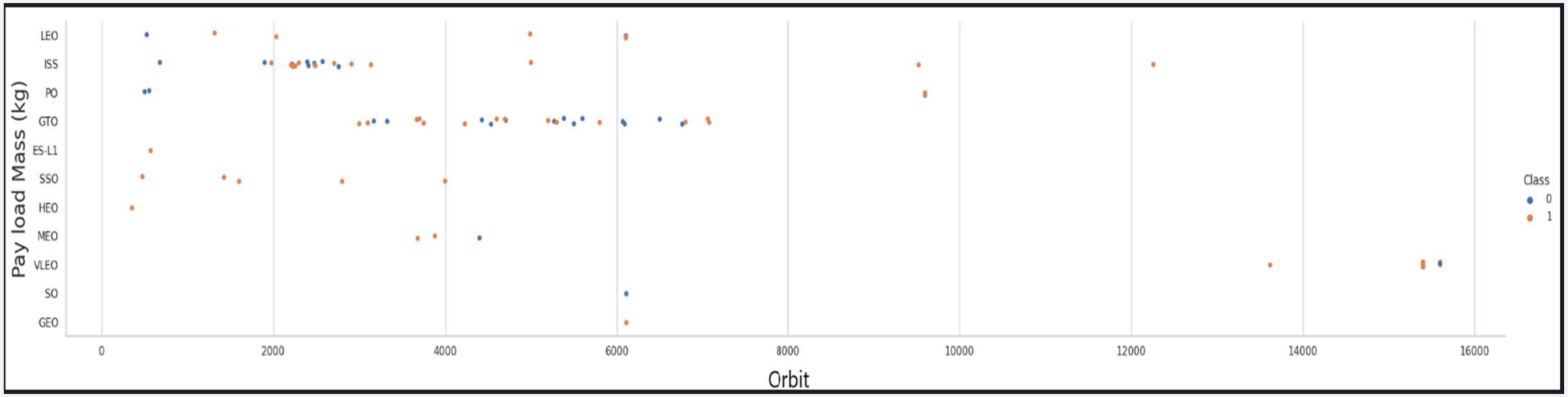
- Show a scatter point of Flight number vs. Orbit type



- In this chart we can have better understanding of how Orbits can influence if the launch is successful or not.
- We observe that SpaceX started with different Orbits and the success rate was fluctuating. In recent launches, the Sun Synchronous Orbit has been the preference and success has improved.

Payload vs. Orbit Type

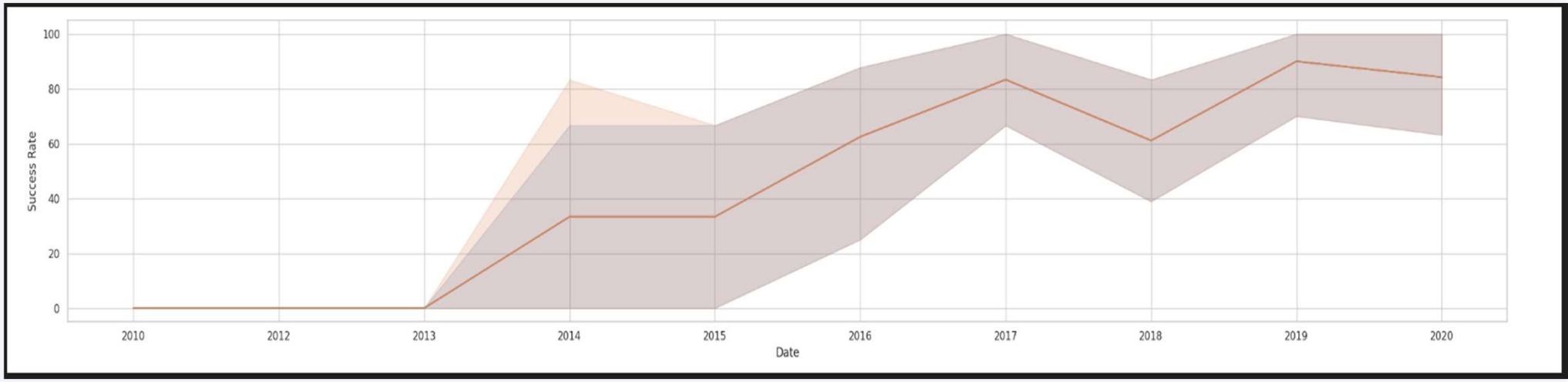
- Show a scatter point of payload vs. orbit type



- There isn't much of a significant correlation between Payload and Orbit.
- We can see that majority of flights are in the GTO (Geosynchronous orbit), which is a very important orbit that monitors many factors of our daily lives such as communication, weather, surveillance, etc

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



- The line chart shows the successful rates along the years. And as we can see the trend after 2013, we show that successful launches are occurring more often, which shows that the effort to make the missions safer, is paying off and we hope to see this trend line reach 100% success rate in the years to come.

All Launch Site Names

- Find the names of the unique launch sites

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- The query returns the unique launch site names.
- We do see CCAFS as SLC-40 and LC-40, which represent the same location, but it seems that the data has wrong data entry, generating this error.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
%%sql
SELECT * FROM SPACEXTBL WHERE (LAUNCH_SITE) LIKE 'CCA%'LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

- Display the first five entries where the launch site name starts ‘CCA’

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596.0

- The value displayed value represents the total payload sent to space where the customer was NASA.
- The information also refers to CRS (Commercial Resupply Services), that indicates that these mission were carrying payloads to the ISS.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) as average FROM SPACEXTBL WHERE booster_version LIKE 'F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
```

average
2534.6666666666665

- An average of how much payload the booster Falcon9 v1 carried on its missions.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%%sql
SELECT Max(Date) as First_Success FROM SPACEXTBL WHERE `Landing_Outcome`='Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

First_Success
22/12/2015

- The date where the first successful launch and landing at a ‘ground pad’ occurred in December 22 of 2015. Which is a very good Christmas gift for all the work done.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select distinct booster_version from SPACEXTBL where Payload_Mass__KG_  Between 4000 and 6000 and LANDING_OUTCOME='Success (drone ship)';
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- The four booster versions were successfully landed at a drone ship and their payloads were bigger than 4000kg and not greater than 6000kg.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT (*) as Outcome FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	Outcome
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- There are many data entries with no information. The correct action to display this, was to first clean all the null entries.
- The query shows us the mission outcome, which points us to an amazing 99% successful outcome.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

- All the boosters displayed above, have carried the maximum payload in their missions.

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
SELECT Landing_Outcome, BOOSTER_VERSION, LAUNCH_SITE, DATE
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)' AND DATE like '%2015';
```

```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Booster_Version	Launch_Site	Date
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	01/10/2015
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	14/04/2015

- The query displays the two failed launches that attempted to land on the drone ship

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%%sql
SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE LANDING_OUTCOME LIKE 'Success' and DATE BETWEEN '04/06/2010' AND '20/03/2017'
-- WHERE LANDING_OUTCOME LIKE 'Success'
GROUP BY LANDING_OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

```
 * sqlite:///my_data1.db
Done.
```

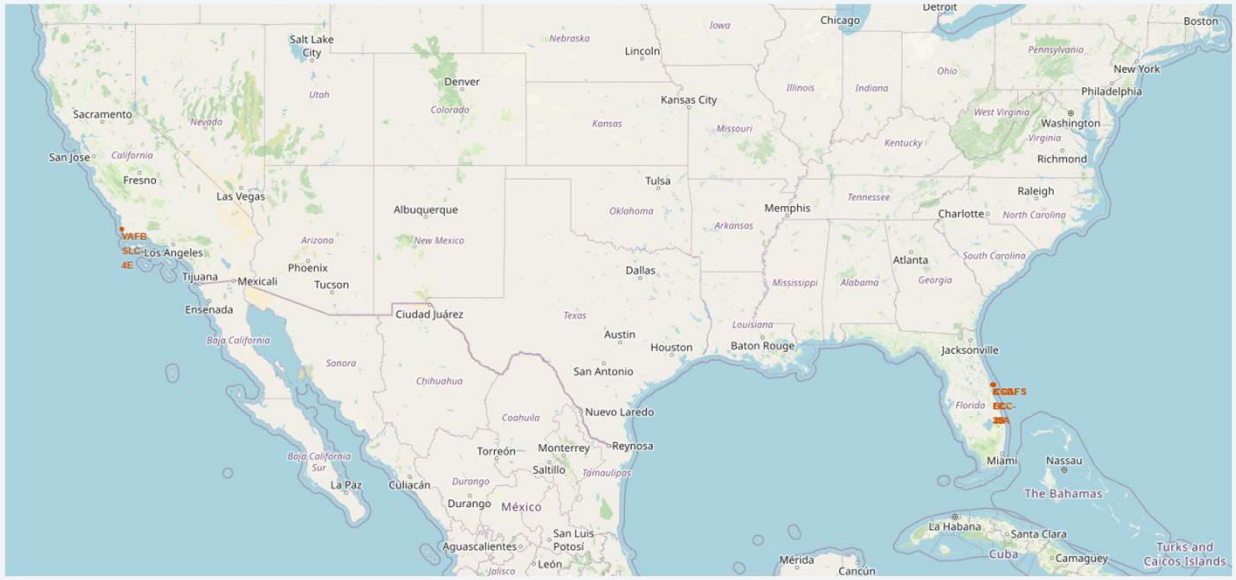
Landing_Outcome	TOTAL_NUMBER
Success	20

- Present your query result with a short explanation here

Section 3

Launch Sites Proximities Analysis

Folium Map: All launch sites

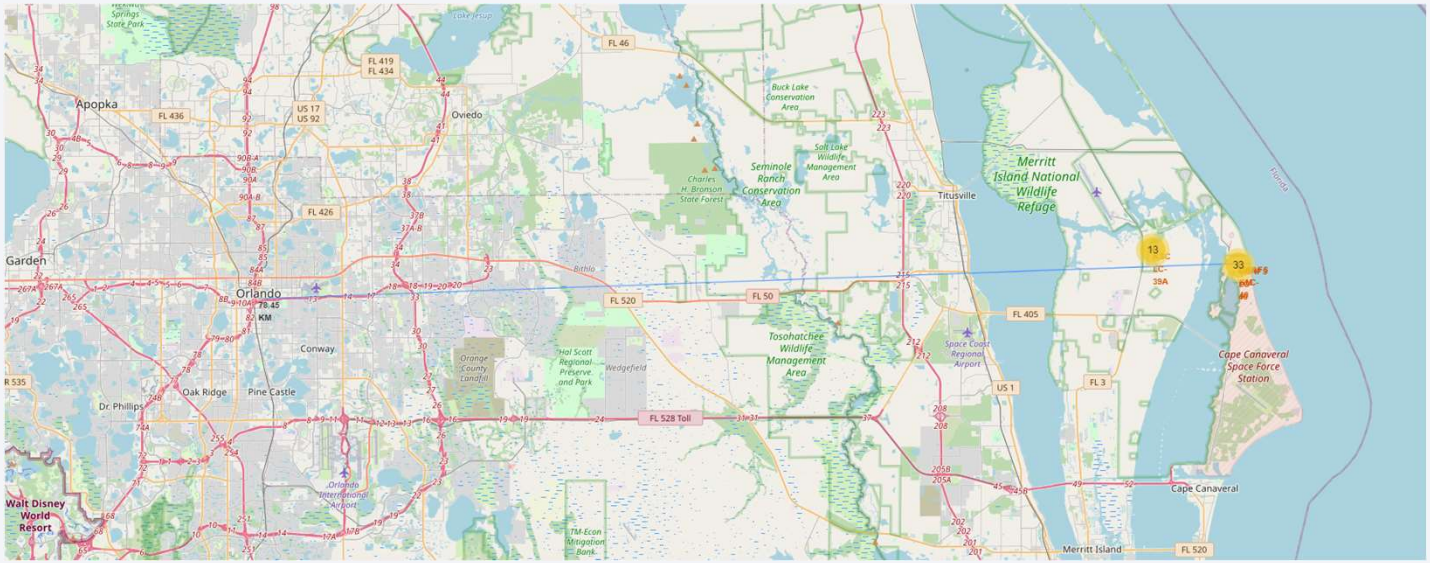


- We observer that the launch sites are located in the USA, however with some similarities, which they all are near the coast, and also in the south of the country, which brings the launches closer to the equator.

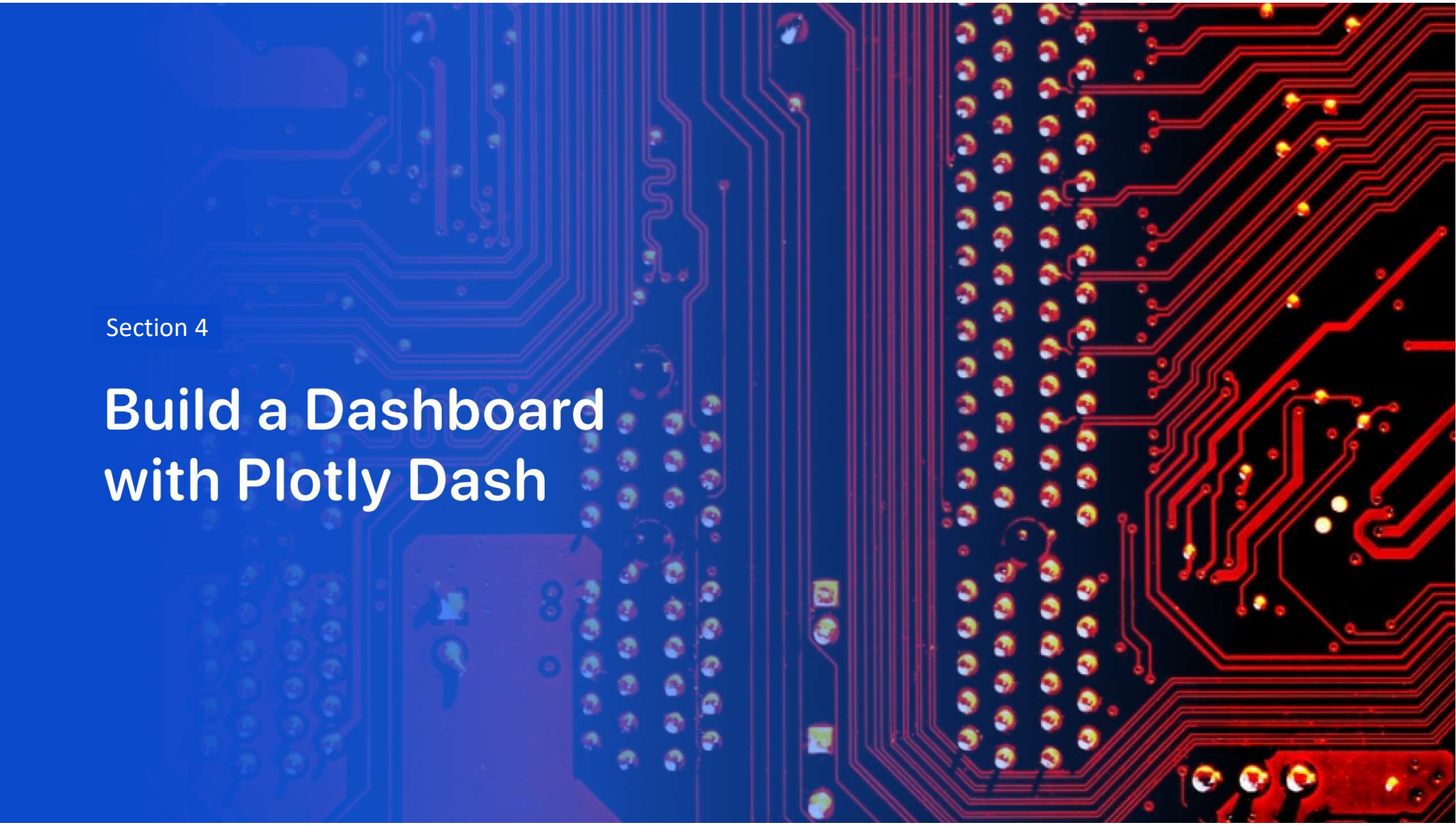
Folium Map: Color Launch Markers

- Both images display the same region of the map.
- The left side, you can observe the number 10 inside the marker. Which shows us that there were 10 launches from this site.
- The right-hand side, we see as colored markers how these launches and landing took place. Green shows us the successful complete mission, as for red the incomplete one.

<Folium Map Screenshot 3>



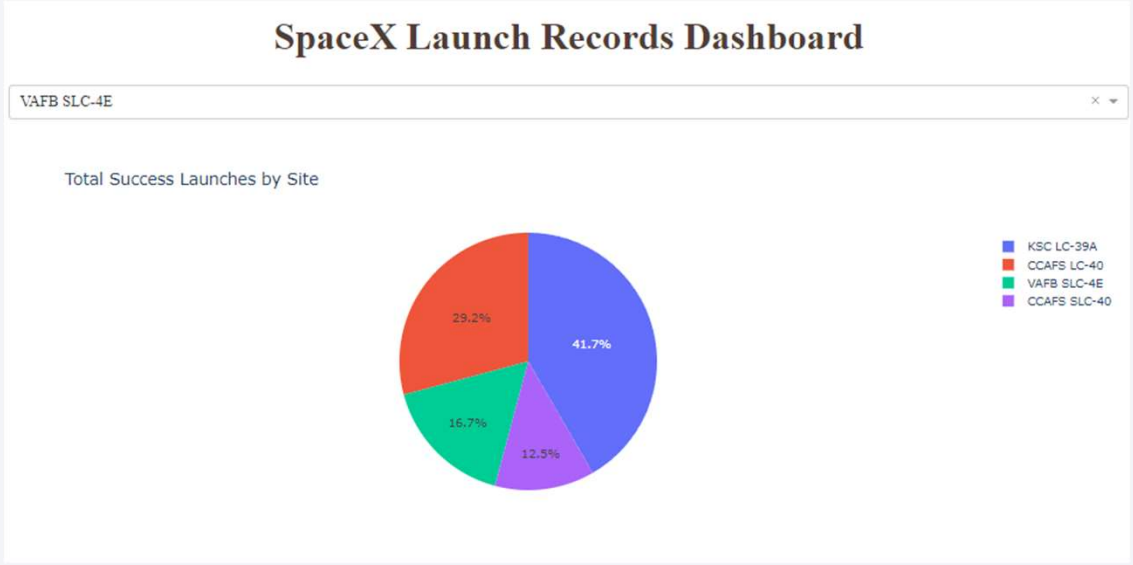
- This is to show the distance from the launch sites located in the coast of Florida to the city of Orlando, which is one of the biggest in the state.
- The distance in a straight line is of 78.45 km. which shows us that the space mission launches are cautiously planned, in order to ensure safety of cities nearby.



Section 4

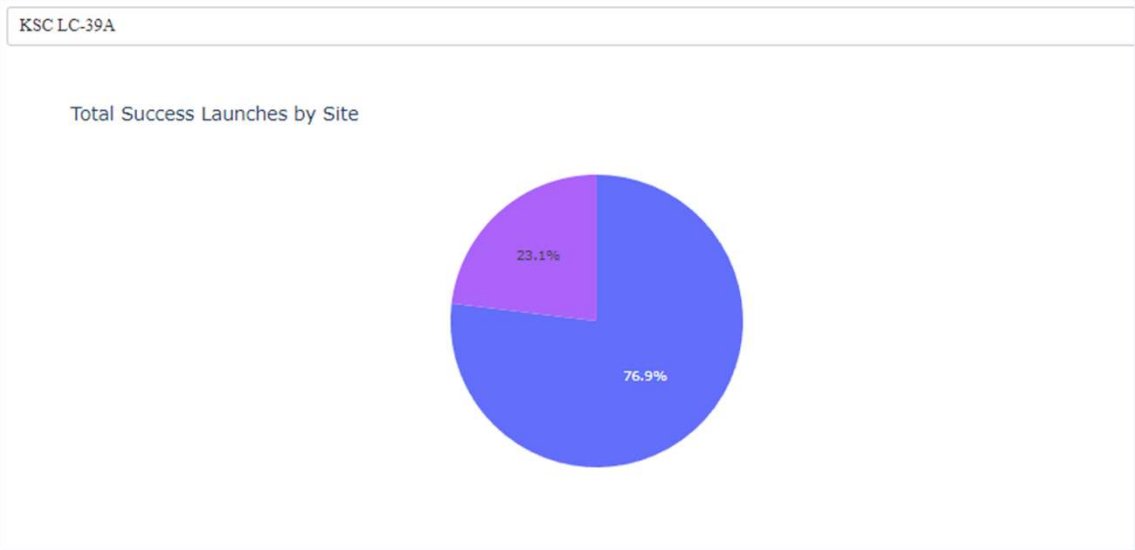
Build a Dashboard with Plotly Dash

Pie Chart – Plotly Dash – Success Rate by Launch Site



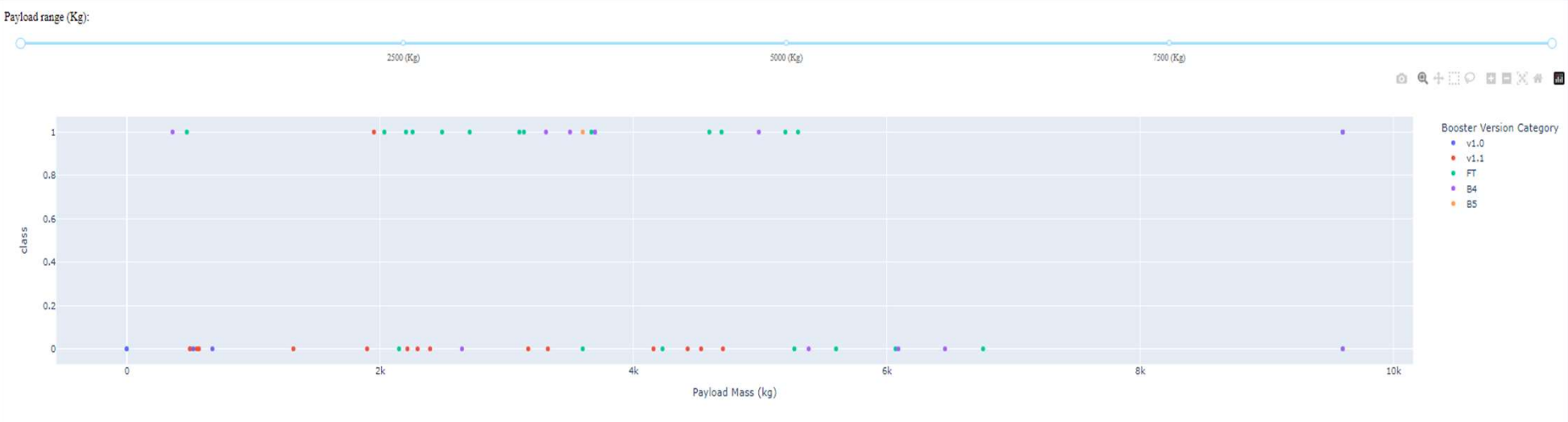
- The chart shows us the success rate of each launch site. And we can observe that KSC has the highest success, and the lowest rate is the CCAFS location.

Pie Chart – Plotly Dash – Success rate for KSC LC-39A



- The pie chart highlights the success rate of the most successful launch site of all.
- We can see that 76.9% of all missions are concluded as planned. As for the remaining 23.1% are all case to study for future improvements.

<Dashboard Screenshot 3>



- The Scatter plot has the success and failure of launches in relation to the payload mass and it has the legend to show us the different versions of boosters used in the missions.
- The Scatter plot done in the Plotly Dash dashboard, provides us with a range selector slider, from which we can choose the payload mass of the missions displayed. Unfortunately, it only shows us up to 10000kg, and as we know, the payload goes all the way up to 15600kg.
- The data also shows 0kg payload mass launches that failed. This could be a data entry error. Further analysis would be ideal.

Section 5

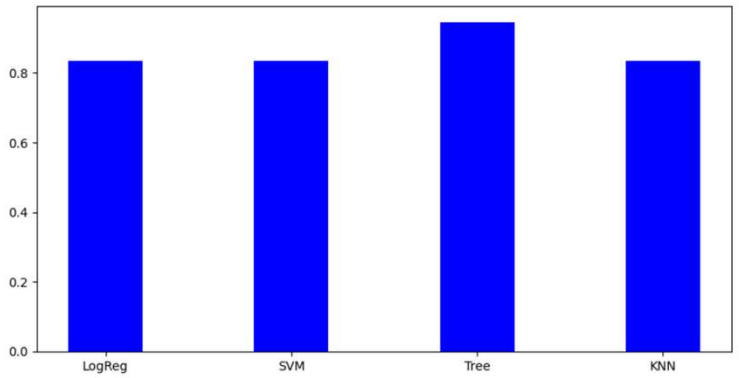
Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

```
Y_bar = [logreg_cv.score(X_test, Y_test), svm_cv.score(X_test, Y_test) , tree_cv.score(X_test, Y_test) , knn_cv.score(X_test, Y_test)]
X_bar = ['LogReg' , 'SVM' , 'Tree' , 'KNN']
```

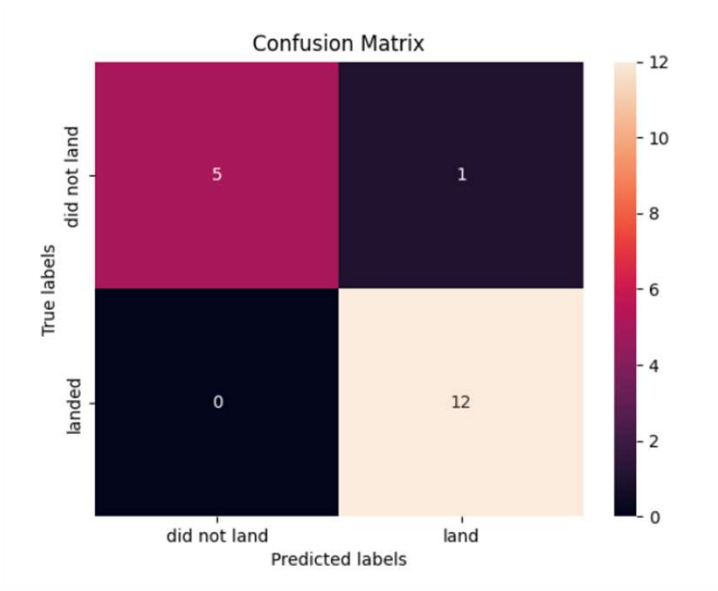
```
fig = plt.figure(figsize = (10, 5))
plt.bar(X_bar, Y_bar, color ='blue', width = 0.4)
plt.show()
```



- As shown, we see that three model predictions have the same accuracy (83.33%) and the Classifier Tree model stands out with a 94.4% accuracy.
- One factor for this interesting result, may be that we don't have sufficient data to have more training sets and test them for a better prediction.

Confusion Matrix

- Show the confusion matrix of the best performing model
- Since all predictive models performed the same, all their confusion matrices are the same.
- The models predicted 12 successful landings when the true label was successful
- The models predicted 5 unsuccessful landings when the true label was unsuccessful landing
- The models predicted 1 success landings when the true label was unsuccessful landings (false positives)
- The model over predicted successful landings.



Conclusions

- In this study, our goal was to develop a predictive model, using machine learning and predictive analysis, to determine if the first stage of rockets will be reused or not.
- The determination of this subject, it allows company such as SpaceX to charge about 100 millions less from its competitors.
- The results we got from our predictions, generated the same results, even though we used four different models. The reason behind this is the amount of data we have available.
- The accuracy of most of our prediction models are of 83.33%. The Tree classifier stands out at a 94.4% Accuracy. A good result, however, we want to be at 100% when we are dealing with rocket launches.
- In order to improve our models and achieve the goal of 100% accuracy, we will need to continue collecting data and make sure that all proper dependent variables are in consideration, so that our independent variable (the successful launch and land) can be predicted safely.
- I'm positive that time and data will take us to our goals.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project
- https://github.com/joshuabls/Applied-Data-Sciece-Capstone
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/labs/module_2/data/Spacex.csv
- https://api.spacexdata.com/v4/launches/past
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Thank you!